DOCUMENT RESUME

ED 343 956 TM 018 166

AUTHOR Dolmans, Diana H. J. M.; And Others
TITLE Assessing Test Validity through the Use of Teachers' Judgments.
PUB DATE Apr 92
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *College Curriculum; Content Analysis; Educational Objectives; Foreign Countries; Higher Education; *Medical Students; *Professors; *Teacher Attitudes; Test Construction; Test Content; Test Items; *Test Validity
IDENTIFIERS Curriculum Based Assessment; Netherlands; *Problem Based Learning

ABSTRACT

A method was developed for assessing the extent to which a test reflects the topics addressed in the problems presented in a problem-based learning environment (curriculum validity). The intended curriculum for a unit from the second year of medical school at the University of Limburg in Maastricht (Netherlands) was analyzed, and an 132-item topic list was constructed to serve as a blueprint of unit content. At the end of the unit, students took an achievement test. The overlap between the intended curricular content, as specified in the topic list, and the information required to answer achievement test items correctly was assessed by four teachers (content specialists). Some curricular content domains were underrepresented on the test or totally absent, and some were overrepresented. It is concluded that the test does not accurately reflect curricular content as specified in the topic list. The method described seems to be a valuable approach for assessing the curricular validity of a test and could be adapted to use in test construction. Two figures and two tables highlight study concepts, and there is a seven-item list of references. (SLD)

Assessing Test Validity Through the Use of Teachers' Judgments[1]

Diana H. J. M. Dolmans, Wim H. Gijselaers and Henk G. Schmidt

University of Limburg

Department of Educational Research

Maastricht, the Netherlands

[1] Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco CA, April 1992.

We gratefully acknowledge the assistence of the teachers.

Correspondence concerning this paper should be addressed to Diana Dolmans, University of Limburg, Department of Educational Research, P.O. Box 616, 6200 MD Maastricht, The Netherlands

Introduction

The ultimate concern in education is student learning. In order to activate student learning effective educational programs are required. Both student learning and program effectiveness are usually measured by achievement tests. Test scores are used as indicators to assess students' progress and to distinguish be ween good and poor students. In addition, test scores are considered as legitimate indicators for policy-oriented purposes. In this case, test scores are used to make policy decisions, such as evaluating program effectiveness for accountability purposes (Airasian & Madaus, 1983). For both applications, knowledge about the extent of overlap between what is tested and what is taught is essential.

If, for example, test scores are used in certifying for graduation, the test needs to represent the intended content domain, otherwise students are examined about subject-matter not included in the unit. Consequently, the test needs to reflect the objectives domain specified for a certain unit. In other words, content validity needs to be assured (Ebel, 1983). However, the test not only needs to be valid against the objectives domain, but also against the schools' curricular materials actually used or against the content actually taught to the students. If a test does not correspond with the curricular materials used or the instruction addressed to the students, then students did not have the opportunity to study the information tested. Students who fail such tests are being penalized for the failures of teachers and not for their own inadequacies.

These considerations, have led to the introduction of two new types of validity by Schmidt, Porter, Schwille, Floden and Freeman (1983), in addition to the widely used concept of content validity: Curricular validity and instructional validity. Content validity asks whether the test accurately reflects the objectives domain specified for development of the test. Curricular validity asks whether the test, established as valid with respect to the domain of objectives, is also consistent with

the curricular materials used in the school system wherein it is to be administered. Instructional validity, on the other hand, is a matter of whether the test, however valid with respect to the objectives, adequately samples the instructional content actually taught to students (Schmidt et al., 1983). If a test has content validity with respect to both objectives and curricular materials, then the objectives domain is likely a proper subset of the curricular materials domain. This is shown in Figure 1. If a test has content validity with respect to both objectives and instructional content, then the objectives domain is a proper subset of the instructional content domain.

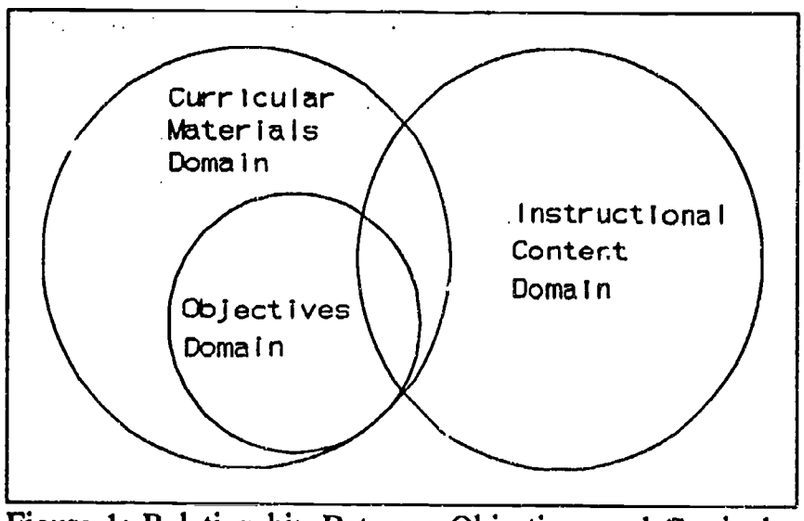

Figure 1: Relationship Between Objectives and Curricular Materials Domain

What type of validity is needed for what type of test is especially relevant in a problem-based learning curriculum. The principal idea behind problem-based learning is that learning should be organized around problems related to the profession, rather than around subjects derived from academic disciplines. Problems usually consist of a description of a set of observable phenomena or events in need of some kind of explanation. Students analyze these problems, attempting to understand the underlying principles or processes through small-group discussion. In doing so, they activate whatever they already know about the problems. However, students' prior knowledge in itself is not sufficient to attain a deep understanding. During discussion usually

questions remain unanswered which subsequently serve as a guide for independent and self-directed learning (Schmidt, 1983).

Consequently, in problem-based learning, students' learning activities are assumed to be dependent on the nature of the problems presented. This implies that the achievement test administered to the students at the end of the unit to measure students' performances should reflect the topics addressed in the problems presented. In other words, the achievement test needs to be valid with respect to both the objectives domain and the curricular materials or the problems. The instructional validity of the achievement test is much more difficult to assess since students in a problem-based curriculum are responsible for their own learning because they largely define themselves the content to be mastered. This self-directedness makes it quite difficult to measure the instructional content addressed to the students in comparison with a teacher-centred approach in which teachers determine what information should be learned, how it is to be learned, and in what sequence. In summary, since problems are the starting point for students' learning activities, the achievement test needs to be valid with respect to both the objectives domain and the problems presented. So, the focus of this paper will be on the latter type of validity, curricular validity.

Determining what type of validity for what type of test seems to be relevant is only one aspect to be considered. Another important facet of test validity involves the technique to be used to assess the extent to which a test is valid. Several approaches have been developed during the last few years to assess test validity. Some of these approaches are solely restricted to the materials covered in textbooks, others attempt to measure effects of instruction. Furthermore, there are considerable differences in the level at which the overlap is estimated such as individual student by item, class by item, individual student by test and class by test. Leinhardt & Seewald (1981) distinguish four major different approaches with respect to test validation. The first approach consists of building a new 'criterion referenced' test

for each new testing-instruction unit. The central focus of this approach is to assess whether a student masters the objectives set for that particular testing-instruction unit in order to decide whether the student can start with the next unit. The mode of analysis for this approach is individual student by test. The second approach tries to ensure test item validity by altering existing tests. Test items are added or removed in line with instructional emphasis. The mode of analysis for this approach is class by item. The third approach makes use of a detailed taxonomy in which it is assumed that the same or similar labels refer to the same content and that different labels refer to different content. Each test item is examined and classified according to a taxonomy to get a visual representation of the areas covered. The test which best mirrors the curriculum is selected. The mode of analysis for this approach is class by item. The fourth approach is based on direct measurement of the overlap between test and instruction by asking judges to estimate the overlap between test and instruction. This fourth approach can be used to calculate instruction-based measures of overlap by asking teachers to estimate the percentage of students at the class level who had been taught the minimum material necessary to pass each test item, class by item level, or by asking teachers to estimate whether a specific student had been taught enough information to answer the item correctly, student by item level. According to Leinhardt and Seewald (1981), the fourth approach can also be used to estimate curriculum-based measures of overlap. These measures are derived from a computer-based curriculum analysis technique in which a dictionary containing information needed to pass an item and a dictionary containing information addressed in curricular materials are compared with each other to estimate the curriculum-based measures of overlap, at the class by item level.

The purpose of this study is to assess curricular validity or the extent to which the test reflects the topics addressed in the problems presented. These curriculum-based measures were estimated by means of content specialists' judgments. This

approach is comparable with the curriculum-based measures described above. The difference is that expert judgments are used in stead of computer-based estimates.

The approach described in this paper consists of the following steps. First the intended curriculum was analyzed. Therefore, a Topic List was constructed consisting of 132 topics covering the unit content as intended by the teachers. These topics refer to students' learning activities expected to be employed during studying the curricular materials. The Topic List can be seen as a blueprint of the unit content or a specification of the curricular materials used, since this list is supposed to cover the intended unit content domains. After the end of the unit an achievement test was administered to the students consisting of 171 test items. The degree of overlap between the intended curricular content, as specified in the Topic List, and the information required to answer the items of the achievement test correctly, was assessed by teachers. These raters had to judge the correspondence between the topics of the Topic List and the items of the achievement test. This correspondence provides insight in the degree to which the intended unit content is tested in the achievement test or the degree to which curricular validity is assured.

The use of content specialists' judgments appears to offer considerable promise as a means for assessing test item validity, according to Rovinelli and Hambleton (1977). The first advantage is that the approach is not dependent on examinee group composition or instructional effects. Second, this approach may not require sophisticated statistical techniques. Third, it is not restricted to highly structured content domains and fourth, it can be implemented easily in practical settings (Rovinelli & Hambleton 1977).

Method

Subjects. This study was conducted at the medical school of the University of Limburg, the Netherlands. The educational program of this medical school is based

on the principles of problem-based learning. The first four years of the problem-based curriculum are structured as a series of six-week units. This study was carried out during the sixth unit of the second year containing 12 problems which are supposed to cover the intended unit content. These problems are related to normal pregnancy, delivery and normal development of children and adolescents.

Instruments. The instruments consisted of a Topic List and an end-of-unit examination. The Topic List contained a list of 132 topics covering the unit content as intended by the teachers. The topics of the Topic List reflected students' learning activities that were expected to be employed during analyzing and discussing these problems. Thus, the topics of the Topic List were derived from the objectives that were intended to be identified during analyzing these 12 problems. Table 1 contains an example of some topics associated with a problem involving objectives related to the normal development of the foetus.

Table 1: An example of some topics

1	Blood circulation of the foetus
2	Exchange of nutrients by the placenta
3	Effects of smoking on body weight of the newborn
4	Oxygen need of the foetus' brains
5	Room of movement in the amniotic fluid

For each topic a Likert-type question was formulated. Teachers were asked to indicate whether each topic is: unimportant (1), fairly unimportant (2), neutral (3), fairly important (4) or important (5), in relation to the intended unit content. The data obtained can be seen as topic-based measures of the intended curriculum.

The other instrument used was an achievement test, including 171 test items of the true-false type. Students' scores on the achievement test consisted of the percentage items correctly answered.

Procedure. Before the beginning of the unit, three teachers were asked to judge the importance of each topic in relation to the unit content in order to validate the Topic List. These ratings were collected before the beginning of the unit to avoid that the flow of the unit could influence teachers' ratings. The achievement test was administered to the students at the end of the unit. A few weeks later, four teachers were asked to judge the correspondence between the Topic List and the content of the test items. Teachers were asked to be raters because they seem to be closest to the implemented curriculum. These raters had to assign the test items to one or more of the topics presented in the Topic List. Of the total of 171 test items initially included in the achievement test, 15 items were removed because of shortcomings in their formulation. Consequently, 156 test items were actually classified. The sequence of the different steps is illustrated below.

Table 2: Steps carried out to assess curricular validity

-1	Construction of the Topic List
-2	Three teachers judged the topics' importance
-3	Beginning of the unit
-4	End of the unit
-5	Achievement test administered to the students
-6	Four teachers judged the overlap between Topic List and achievement test

Analysis. The average importance of the total Topic List was 3.53 (*SD*=.87) indicating that most topics were judged neutral (3) or fairly important (4) in relation to the unit content. Out of these 132 topics, only 19 topics had an average importance

below neutral (3). These topics were removed from the Topic List because they appeared to be fairly unimportant (2) or unimportant (1) in relation to the unit content. Consequently, the Topic List contained 113 topics which were judged to be valid in relation to the intended unit content.

A test item was judged as corresponding with the curricular content as specified in the Topic List, if at least three out of four raters matched a similar topic to a test item. So, the cut-off score to separate 'valid' from 'non-valid' test items was .75.

Results

The overlap between the Topic List and the achievement test illustrates the extent to which the test's items reflect he intended unit content or the test's curricular validity. Three out of four raters agreed that 106 out of the 156 test items could be assigned to one or more topics of the Topic List. In other words, the raters agreed that 67.9 percent of the test items reflected the curricular materials as specified in the Topic List. The other 32.1 percent of the achievement test's items consisted of 25 percent about which the raters did not agree and 7.1 percent (32.1 - 25) which did not reflect the curricular materials.

The number of topics covered by the 106 test items were 43. In total 43 out of 113 topics (38 percent) of the Topic List were reflected by the content of the test items. In summary, 38 percent of the topics in the Topic List were tested by 68 percent of the items of the achievement test. Consequently, less than one half of the intended curricular content got tested by the achievement test according to the authors' methodology.

In order to assess the test's curricular validity, not only the overlap between the achievement test and the Topic List needs to be estimated. Also the degree to which the test items are evenly balanced across the intended subject-matter is of importance. Since this study was conducted in a problem-based curriculum in which

the curricular materials were organized around 12 problems, the topics could also be classified into the 12 problems. Moreover, since the raters assigned the test items to the topics, it was also possible to calculate the number of items included in the achievement test for each problem. This number varies between 5 and 29. The average importance of the topics for each problem or the average topic-based measures of the intended curriculum did not differ across problems ($F(11,375)=1.45$, $p=.15$). Consequently, these teachers considered each problem to be of equal importance. From this point of view it is expected that the test items are also equally distributed among the 12 problems. Figure 2 contains a summary of the number of test items for each problem.

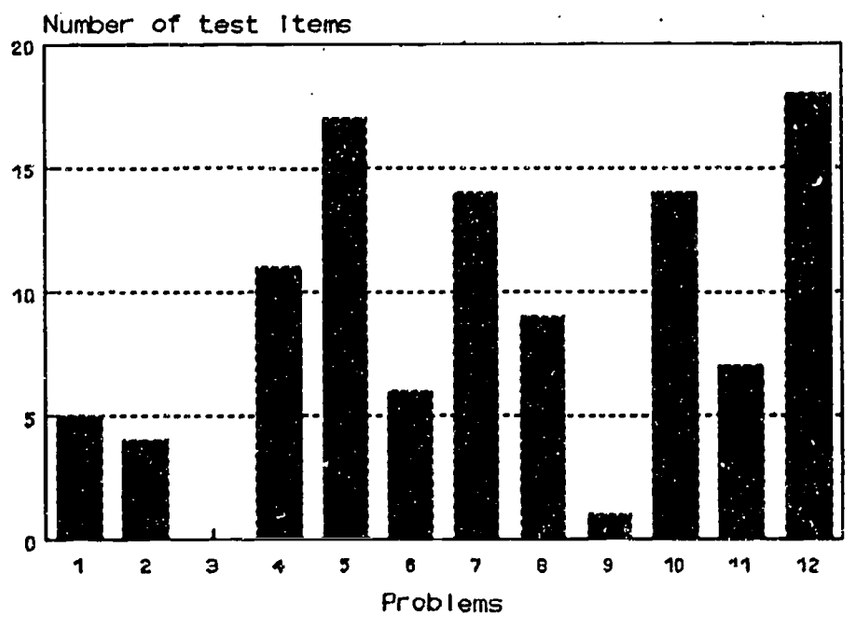

Figure 2: Number of test items for each problem

The number of test items for each problem vary between 0 and 18 as can be seen in Figure 2. Problem 3 contains zero items. Problem 1, 2 and 9 contain relatively less items. The small number of items for problem 1 is astonishing because this problem consisted of two parts that were analyzed during two meetings, as opposed to the

other problems which each were analyzed during one meeting. The number of items included in the achievement test concerning problem 5 and 12 were relatively large.

In summary, it can be concluded that the items of the achievement test are not evenly balanced across the 12 problems and with respect to one problem not a single item was included. These results indicate that, according to the authors' methodology, the achievement test does not adequately cover the curricular materials.

Conclusion and discussion

The purpose of this study was to present a method to assess curricular validity by means of content specialists' judgments about the correspondence between curricular content and test content. The curricular content was specified in a list of topics covering the 12 problems considered to represent the curricular materials. Average problems' topic-based measures of the intended curriculum revealed that each problem was judged to be of equal importance in relation to the course content. These results would suggest that the test items included in the achievement test were also equally distributed among the problems. The results of Figure 2, however, seem to indicate shortcomings in the achievement test. Some curricular content domains are under-represented in the test or totally absent and some domains are over-represented. Topics which were not covered in the test were for the most part related to pedagogical aspects of the development of children and to social and cultural aspects of pregnancy. On the other hand, some topics were over-represented, such as psychosexual development and topics related to the medical examination of the new-born. In summary, it can be concluded that the test does not accurately reflect curricular content as specified in the Topic List. This implies that valid inferences about student learning or program effectiveness can not easily be drawn. According to the methodology used in this study, the achievement test does not seem very

applicable for purposes of assessing student learning or measuring program effectiveness, because the test does not adequately cover curricular content.

Alternative explanations for the rather low percentage of overlap between the curricular materials and the test content may however exist. One of the problems in defining a domain is the level of detail to be contained. The low percentage of overlap may be due to differences in the degree of specificity in formulation between test items and topics. The topics might be formulated more broadly than the test items, since raters assigned only 43 topics to 106 test items. This makes it possible that the raters mainly selected the most broadly formulated topics of the Topic List. Second, the number of topics selected by the raters may depend on the degree to which the topics of the Topic List are mutually exclusive. Consequently, the overlap is contingent upon the accuracy of estimation which is narrowly related to the complex nature of the subject-matter domain (Leinhardt, 1983). In summary, the results presented may depend on the closeness of the match procedure.

The use of content specialists' judgments has several advantages, such as the independence on examinee group composition and the independence on sophisticated statistical techniques. Furthermore, Rovinelli and Hambleton (1977) assumed that measuring test validity by making use of judgments is not restricted to highly structured content domains. The method presented, on the other hand, seems to be slightly restricted to highly structured content domains because the unit content needs to be divided into topics in order to construct a Topic List.

This study's aim was to measure test validity through the use of teachers' judgments. As such, the method described is an a posteriori approach which is used after the test items are written. In general, test validity needs to be reached during the construction phase of a test. The method presented can also be applied during the *development* of a test. A Topic List containing the intended unit content needs to be constructed which can serve as a blueprint for the achievement test. In order to ensure test item validity a direct relationship between a topic and an item should be developed during

a test's construction phase. In this a priori approach also teachers' judgments are used to ensure test item validity. In summary, the method described seems to offer a valuable approach as a means for assessing test validity and can easily be implemented in practical settings.

References

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: policy issues. *Journal of Educational Measurement, 20, 2,* 103-118.

Ebel, R. L. (1983). The Practical Validation of Tests of Ability. *Educational Measurement: Issues and Practice,* Summer, 7-10.

Leinhardt, G. (1983). Overlap: Testing Whether It Is Taught. In: G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 153-169). Hingham, M.A.: Kluwer-Nijhof Publishing, Boston.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18, 2,* 85-95.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Validity. *Tijdschrift voor Onderwijsresearch 2, 2,* 49-60.

Schmidt, H.G. (1983). Problem-based learning: rationale and description. *Medical Education, 17,* 11-16.

Schmidt, W. H., Porter, A. C., Schwille, J. R., Floden, R. E., & Freeman, D. J. (1983). Validity as a Variable: Can the Same Certification Test Be Valid for All Students. In: G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 133-151). Hingham, M.A.: Kluwer-Nijhof Publishing, Boston.